

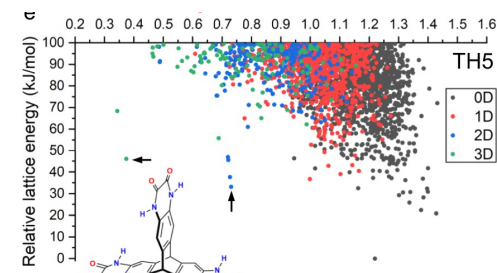
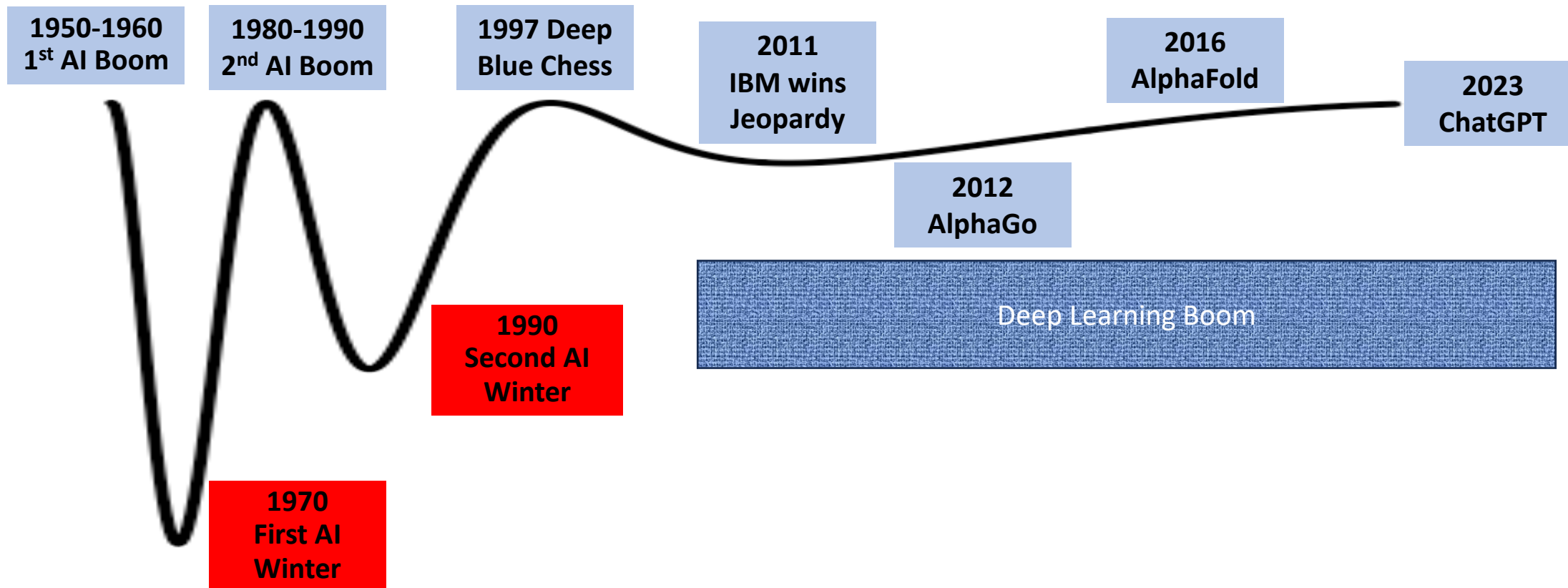
Will an AI win a  
chemistry Nobel Prize  
and replace us?



**PSDI**  
PHYSICAL SCIENCES  
DATA INFRASTRUCTURE

Simon Coles & Jeremy Frey  
School of Chemistry  
University of Southampton

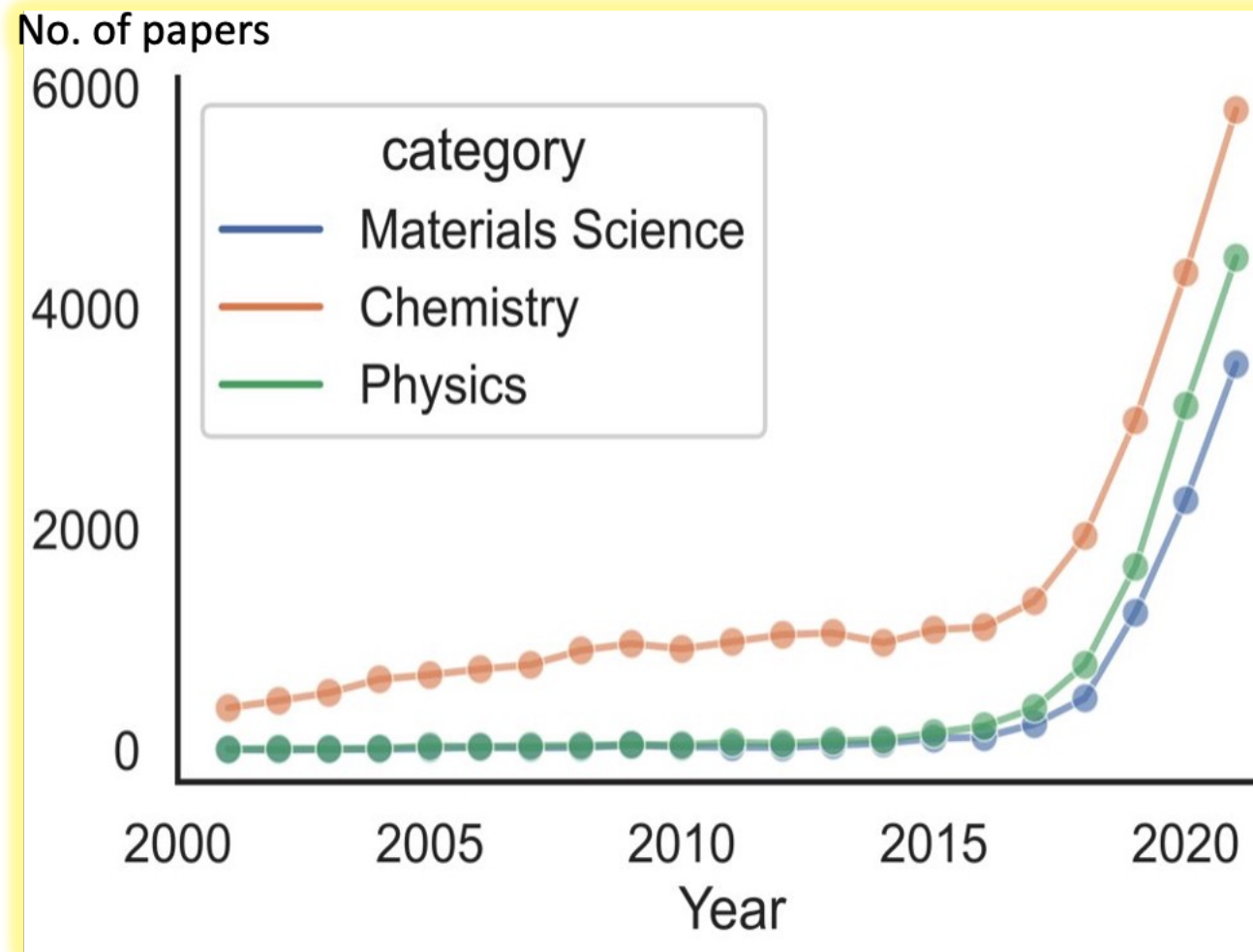
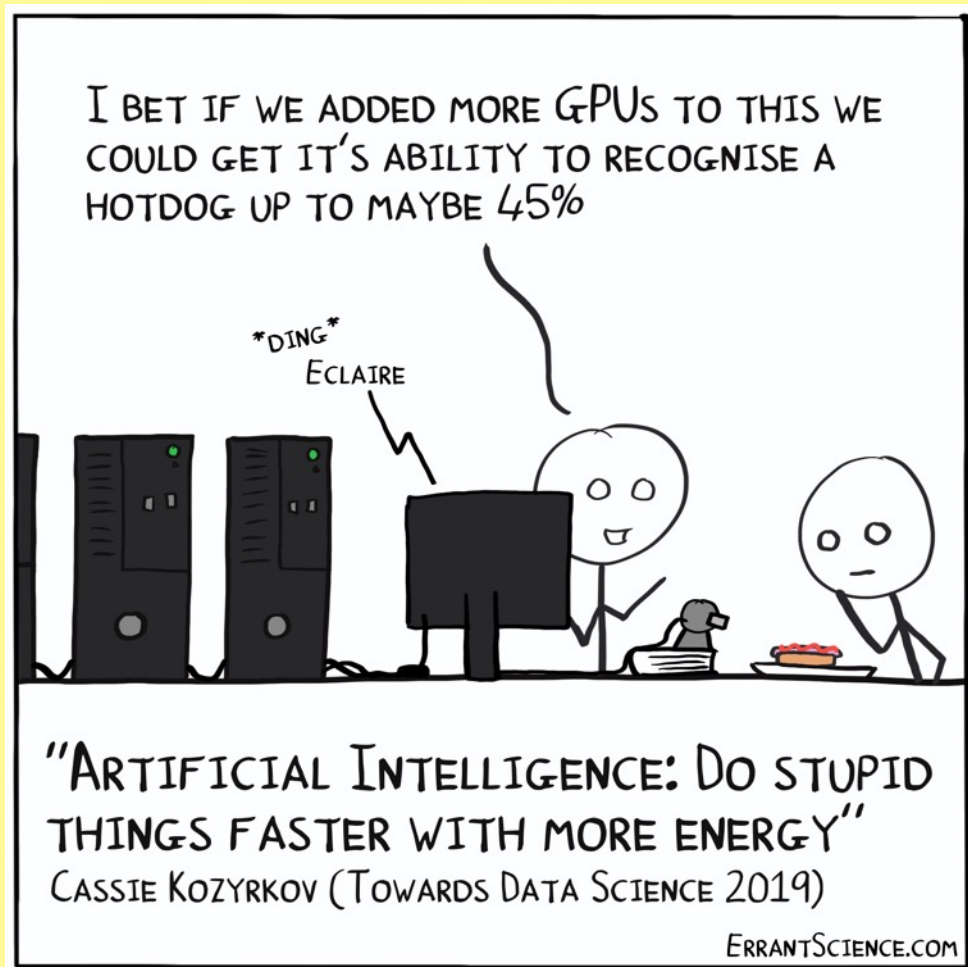
[@colessj](mailto:@colessj) 0000-0001-8414-9272



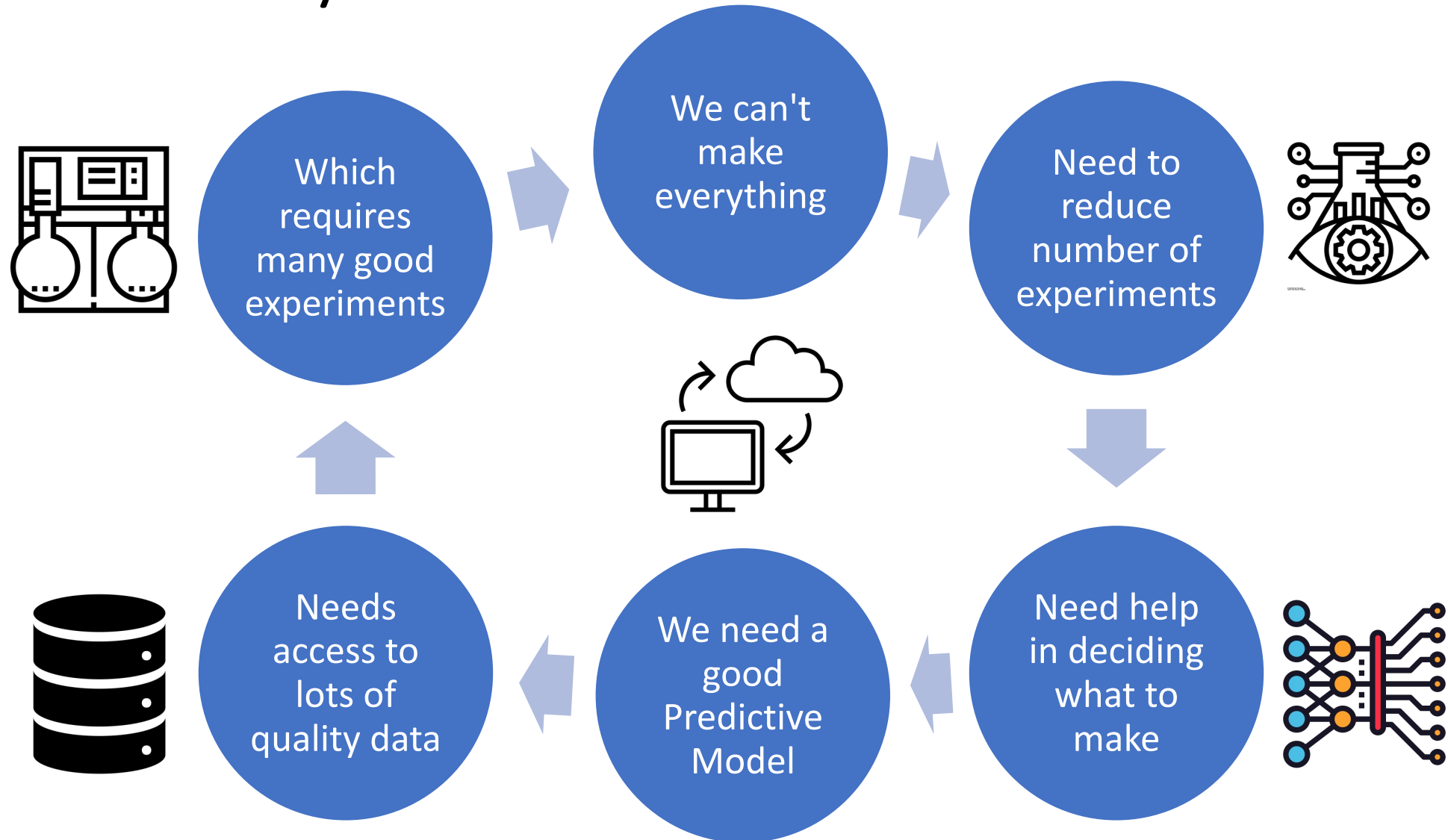
<https://doi.org/10.1038/s41467-021-21091-w>



# Augmented Chemical Intelligence?



# Need to be more insightful, creative, efficient, environmentally conscious...



# From QSAR to Machine Learning

Relevant Data	Equation / Model	Name
Known	Known	Theory
Unknown	Known	Unproved theory
Known	Unknown	Statistical Modelling (QSAR)
Unknown	Unknown	Machine Learning



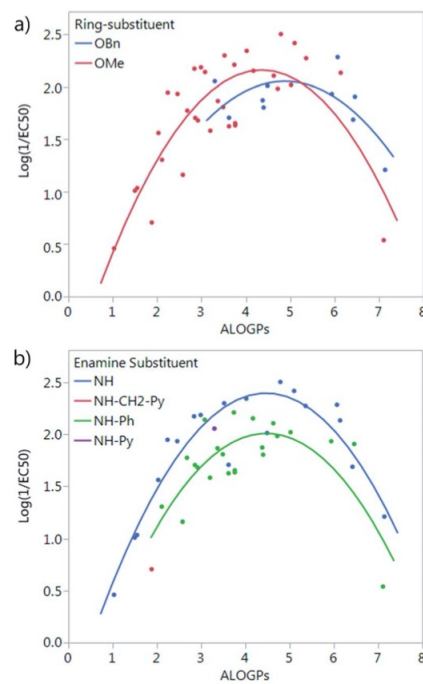
View Article Online  
View Journal | View Issue

## QSAR analysis of substituent effects on tamblamine anion transporters††

Nicola J. Knight,<sup>a</sup> Elsa Hernandez,<sup>b</sup> Cally J. E. Haynes,<sup>§</sup> Nathalie Busschaert,<sup>¶</sup> Harriet J. Clarke,<sup>a</sup> Koji Takimoto,<sup>c</sup> Maria Garcia-Valverde,<sup>b</sup> Jeremy G. Frey,<sup>\*,a</sup> Roberto Quesada<sup>\*,a</sup> and Philip A. Gale<sup>\*,a</sup>

The transmembrane anion transport activity of 43 synthetic molecules based on the structure of marine alkaloid tamblamine were assessed in model phospholipid (POPC) liposomes. The anionophoric activity of these molecules showed a parabolic dependence with lipophilicity, with an optimum range for transport efficiency. Using a quantitative structure-transport activity (QSAR) approach it was possible to rationalize these results and to quantify the contribution of lipophilicity to the transport activity of these derivatives. While the optimal value of log *P* and the curvature of the parabolic dependence is a property of the membrane (and so similar for the different series of substituents) we found that for relatively simple substituents in certain locations on the tamblamine core, hydrophobic interactions clearly dominate, but for others, more specific interactions are present that change the position of the membrane hydrophobicity parabolic envelope.

Received 16th October 2015  
Accepted 27th November 2015  
DOI: 10.1039/C5SC03932K  
www.rsc.org/chemicalscience



## Machine Learning: Data-driven Modelling

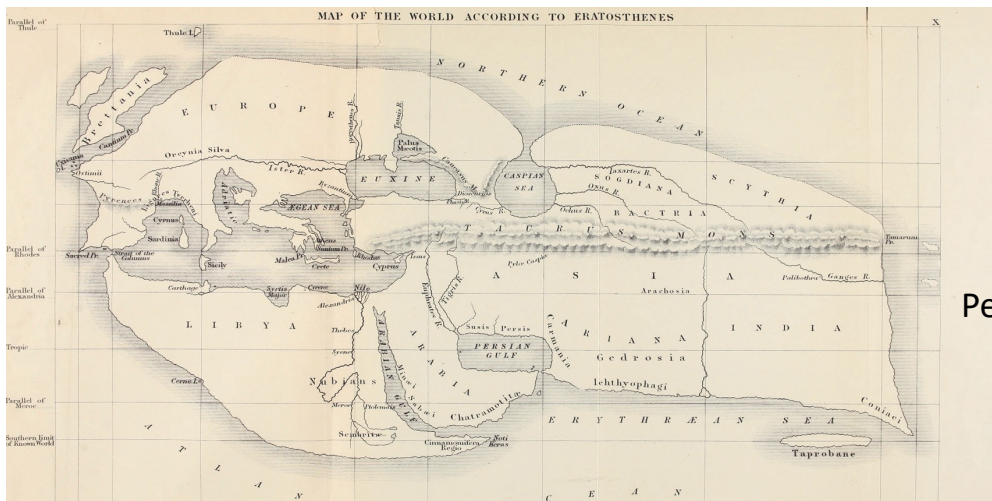
Data	$\{\mathbf{x}_n, \mathbf{t}_n\}_{n=1}^N$ $\{\mathbf{x}_n\}_{n=1}^N$
Function Approximator	$\mathbf{t} = f(\mathbf{x}, \boldsymbol{\theta}) + \nu$
Parameter Estimation	$E_0 = \sum_{n=1}^N \{ \ \mathbf{t}_n - f(\mathbf{x}_n; \boldsymbol{\theta})\  \}^2$
Prediction	$\hat{\mathbf{t}}_{N+1} = f(\mathbf{x}_{N+1}, \hat{\boldsymbol{\theta}})$
Regularization	$E_1 = \sum_{n=1}^N \{ \ \mathbf{t}_n - f(\mathbf{x}_n)\  \}^2 + r(\ \boldsymbol{\theta}\ )$
Modelling Uncertainty	$p(\boldsymbol{\theta}   \{\mathbf{x}_n, \mathbf{t}_n\}_{n=1}^N)$
Probabilistic Inference	$E[g(\boldsymbol{\theta})] = \int g(\boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta} = \frac{1}{N_s} \sum_{n=1}^{N_s} g(\boldsymbol{\theta}^{(n)})$
Sequential Estimation	$\boldsymbol{\theta}(n-1 n-1) \longrightarrow \boldsymbol{\theta}(n n-1) \longrightarrow \boldsymbol{\theta}(n n)$ Kalman & Particle Filters; Reinforcement Learning





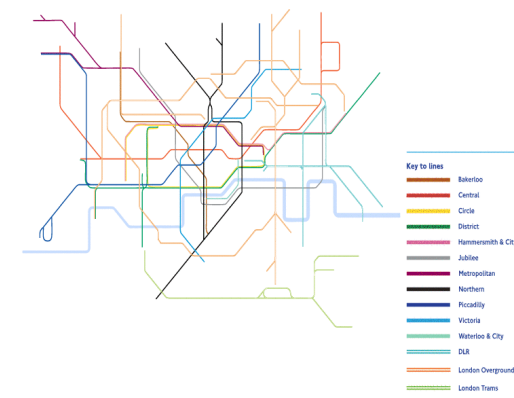
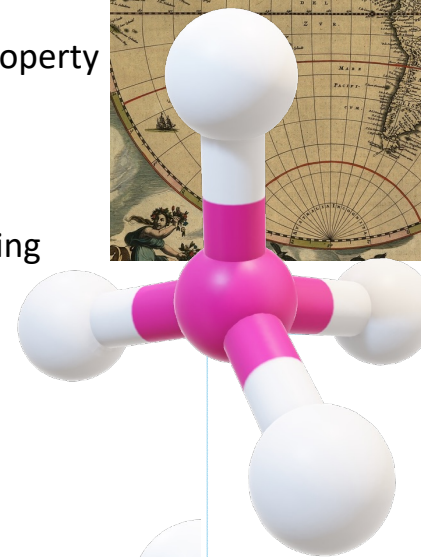
# The need to map chemical space?

The First Law of Geography:  
"Everything is related to everything else,  
but near things are more related than  
distant things".



- Not just big, high dimensional (need to consider Chemical Space-Time)
- What do we mean by related?
- What do we mean by near?
- In high dimensions 'distance' is less well defined

Structure  
Characterisation  
Performance  
Property  
Processing  
Dimension  
Data  
Diversity  
Shape  
Complexity

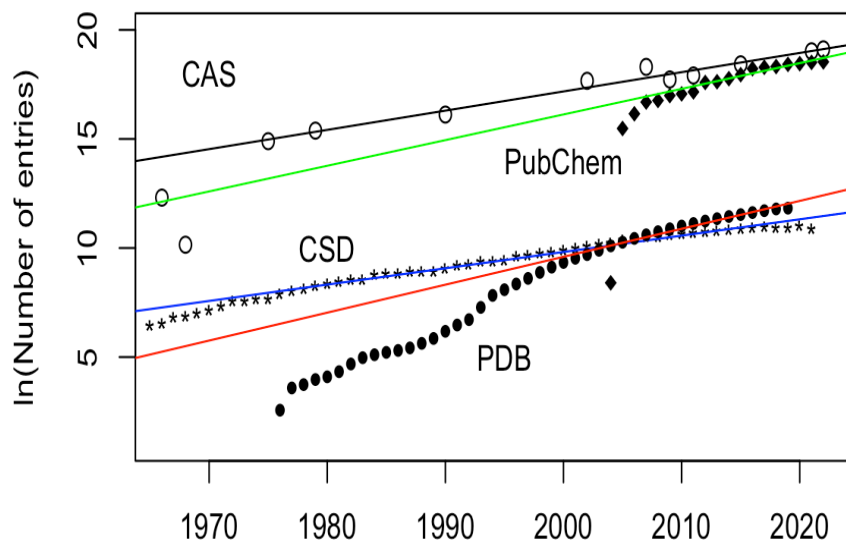


u/a\_wandering\_chemist

# Data, Data everywhere - but not enough to model?

Some collections of trusted, curated data

- Rapid increase in the number of *available* crystal structures
- By no means capturing all structures (being) determined

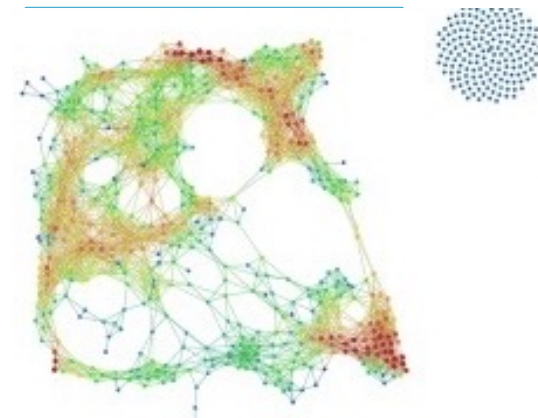
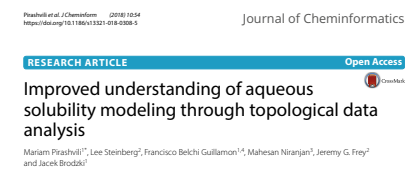


By far, the majority of data is *unstructured*

- Residing in articles, theses, patents (not particularly *accessible*)
- In difficult to process formats
- Never left the lab book...

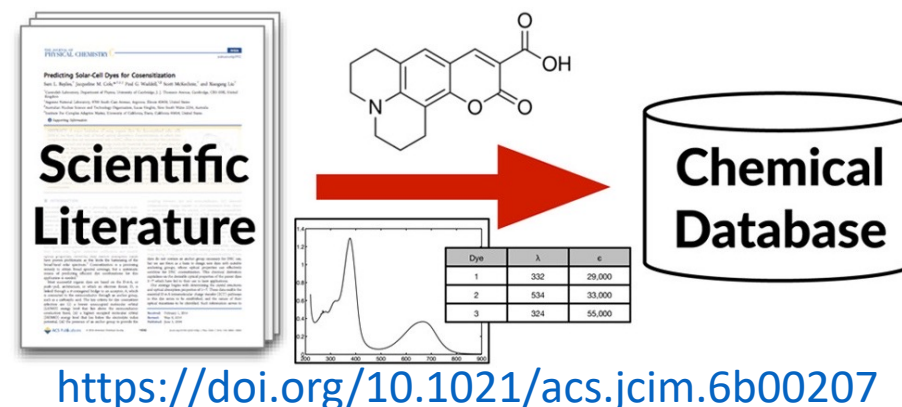


- Beginning to move towards repositories, metadata standards, descriptors
- Need AI suitable ways to structure data e.g. Topology (shape of data)



# Culture change required e.g. compiling non-curated/unstructured data

- Generally only published, successful outcome, data is made available
- Extracting data is time consuming...

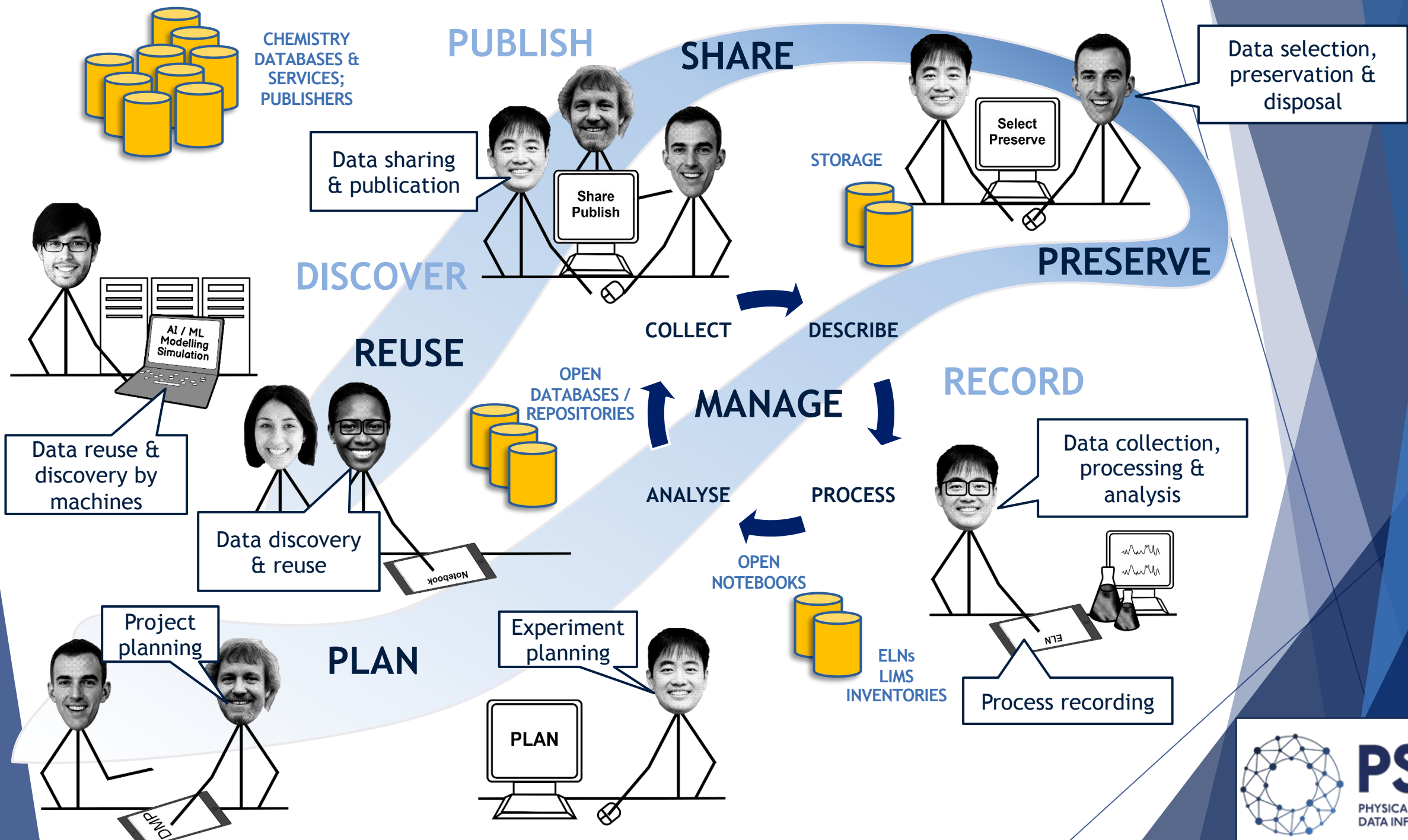


- It takes as much time to convert data to make it usable
- A clear need for data standards...

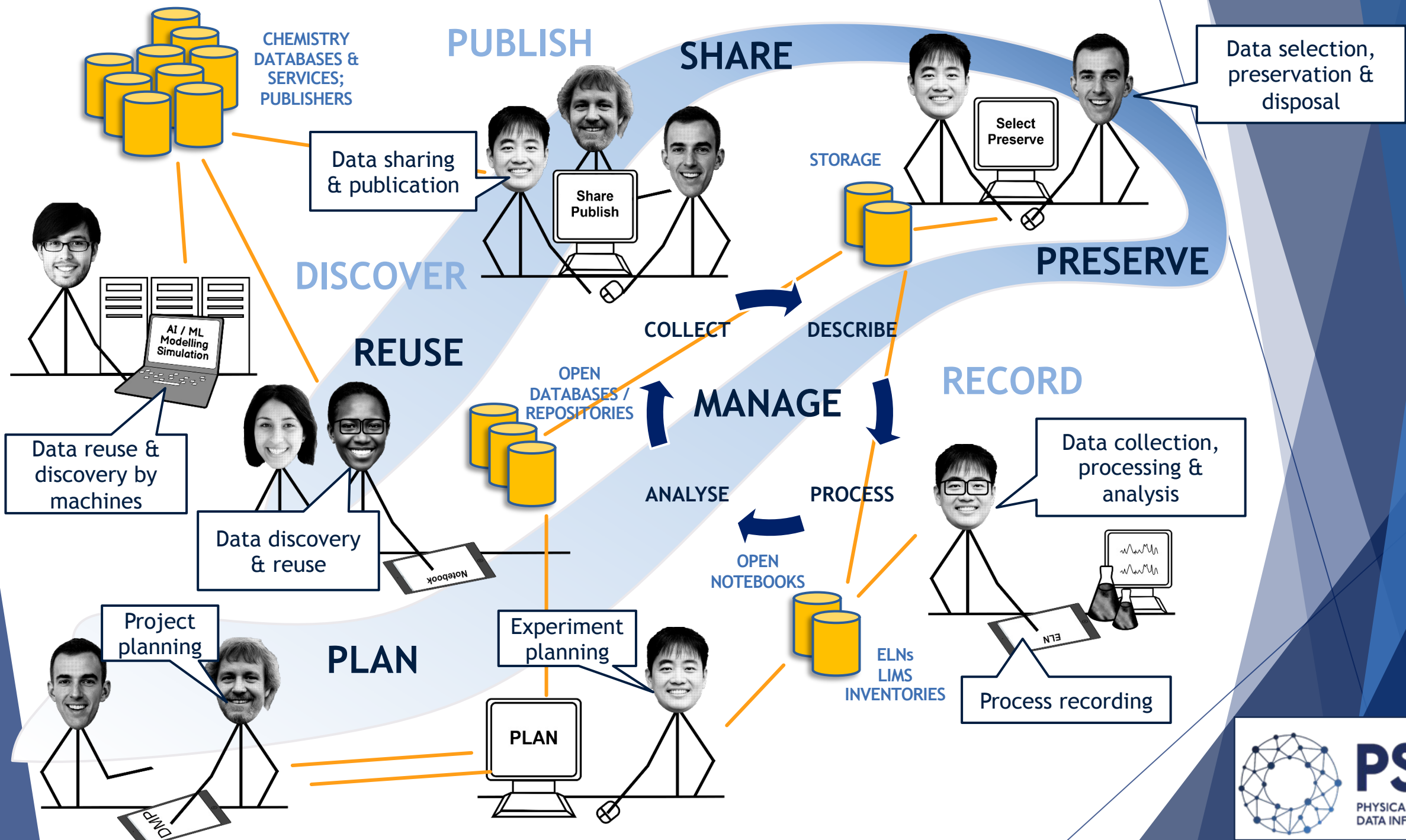


**Desperate need for purpose-built, formal data infrastructures**





**PSDI**  
PHYSICAL SCIENCES  
DATA INFRASTRUCTURE



# We are still in a liminal period!

A key concern:

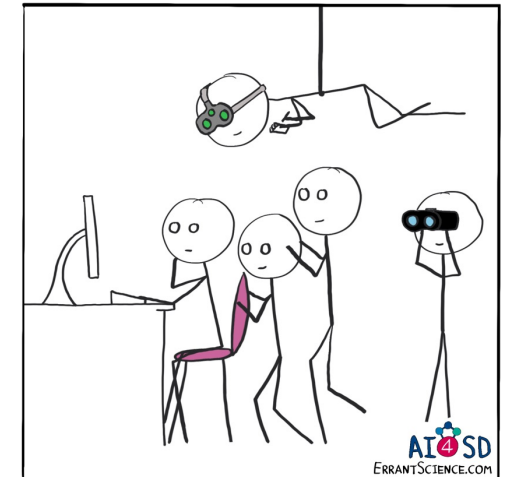
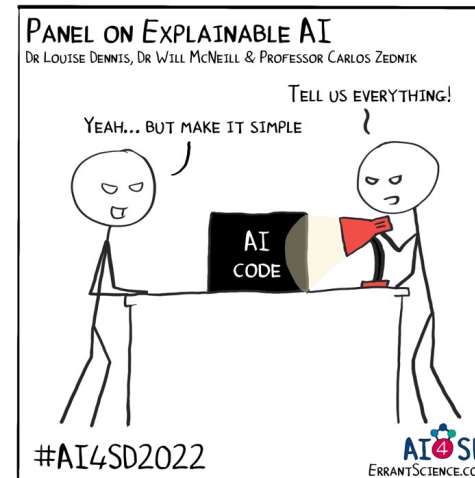
- e.g. Are GPS & Mapping leading to decline in spatial awareness?
- Overreliance on AI may put us in intellectual debt



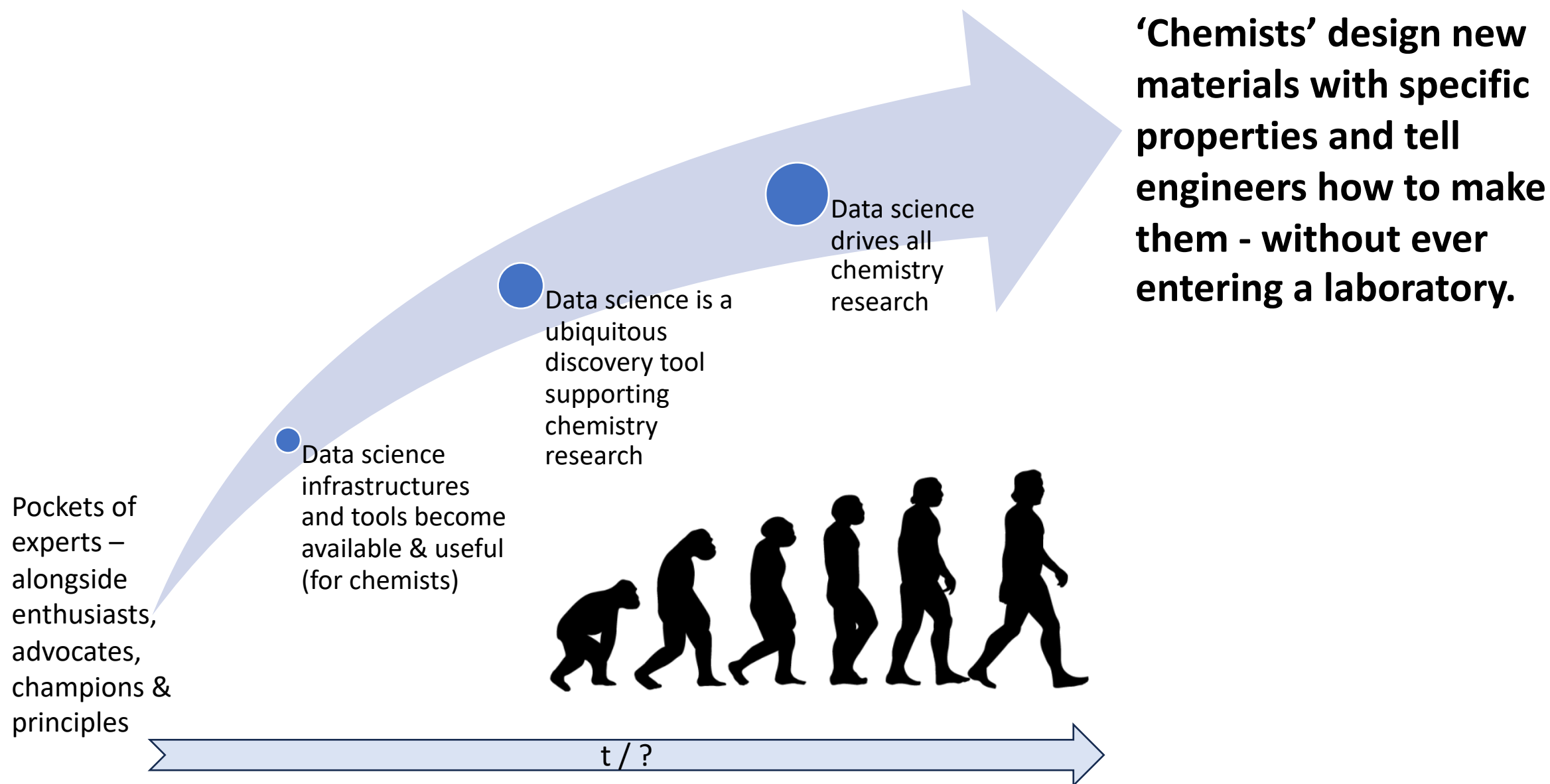
## Ethical & Explainable AI necessary for scientific discovery

The need for trustworthiness

- Provenance of data / training sets
- Benchmark data
- Ability to scrutinise / understand models and methods



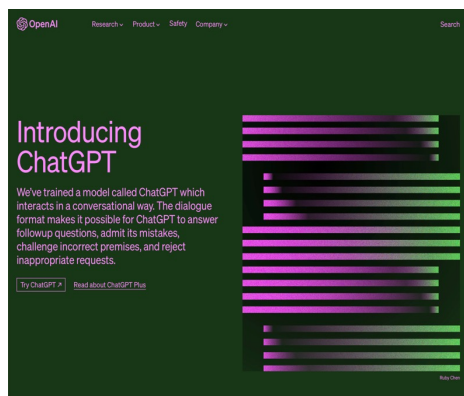
# Evolution of digital chemistry data





# Will an AI win the Nobel Prize and replace us?!

“As an AI language model, I cannot predict the future... However, it is possible for an AI to contribute significantly to chemistry research that could be awarded a Nobel Prize”



”However, it is important to note that the Nobel Prize is awarded to individuals or groups of individuals, not to machines or algorithms.... Even if AI plays a critical role... the prize would likely be awarded to the human scientists who developed and applied the AI methods.”

“...Do you know who will find these things out? Not our AI and ML systems, although I’m sure they’ll help whenever possible. No, it is going to be us. Just like it always has been. The law of conservation of data...”



Derek Lowe

**In the Pipeline  
Chemistry World**

